Full length article

Uncertainty-driven ensembles of multi-scale deep architectures for image classification

Juan E. Arco [a,b,c,*], Andrés Ortiz [b,c], Javier Ramírez [a,c], Francisco J. Martínez-Murcia [a,c], Yu-Dong Zhang [d], Juan M. Górriz [a,c]

[a] *Department of Signal Theory, Networking and Communications, University of Granada, 18010 Granada, Spain*
[b] *Department of Communications Engineering, University of Malaga, 29010 Malaga, Spain*
[c] *Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), Spain*
[d] *School of Informatics, University of Leicester, Leicester, LE1 7RH, UK*

ARTICLE INFO

ABSTRACT

The use of automatic systems for medical image classification has revolutionized the diagnosis of a high number of diseases. These alternatives, which are usually based on artificial intelligence (AI), provide a helpful tool for clinicians, eliminating the inter and intra-observer variability that the diagnostic process entails. Convolutional Neural Network (CNNs) have proved to be an excellent option for this purpose, demonstrating a large performance in a wide range of contexts. However, it is also extremely important to quantify the reliability of the model's predictions in order to guarantee the confidence in the classification. In this work, we propose a multi-level ensemble classification system based on a Bayesian Deep Learning approach in order to maximize performance while providing the uncertainty of each classification decision. This tool combines the information extracted from different architectures by weighting their results according to the uncertainty of their predictions. Performance is evaluated in a wide range of real scenarios: in the first one, the aim is to differentiate between different pulmonary pathologies: controls *vs* bacterial pneumonia *vs* viral pneumonia. A two-level decision tree is employed to divide the 3-class classification into two binary classifications, yielding an accuracy of 98.19%. In the second context, performance is assessed for the diagnosis of Parkinson's disease, leading to an accuracy of 95.31%. The reduced preprocessing needed for obtaining this high performance, in addition to the information provided about the reliability of the predictions evidence the applicability of the system to be used as an aid for clinicians.

1. Introduction

Nowadays, the use of medical images has revolutionized the diagnosis of a high number of diseases. The reduction in the cost during the acquisition of the images and the improvement in their quality have popularized their use in the diagnostic process. However, this is not a straightforward task and success in the detection of a certain pathology depends on many factors. One of the most important is that diagnosis is still largely dependent on the expertise of the clinician. The patient can have symptoms that are compatible with a high number of diseases. Besides, some anatomical structures present in the images can be similar to other abnormal findings associated with a disease, affecting the expert's opinion. This leads to a manual, time-consuming process that has inter and intra-observer variability, which may delay the diagnosis and the election of the treatment. To mitigate this issue, the use of computer aided diagnosis (CAD) systems based on artificial

intelligence (AI) can play a decisive role. These methods rely on the idea that they can learn the main features from a group of samples. Then, the algorithm is able to generalize this knowledge, identifying these features when new samples are analyzed [1–5].

Following this framework, previous works have employed machine learning (ML) algorithms for the automatic detection of a wide range of pathologies such as Parkinson's or Alzheimer's disease [6–9], and most recently, pneumonia [10–14]. Classification systems employed in CAD tools have the following general structure: (i) delimitation of the regions of interest (ROI) to focus the analysis on them, (ii) features extraction from these regions, (iii) classification based on those features. Unlike classical methods based on the extraction of predefined features, deep neural networks build a specific feature space for the optimal class separation by means of a learning process. The emergence of these approaches has revolutionized the automatic classification of

---

medical images [15–22]. It is clear that deep learning models can effectively identify the presence of a certain pathology. However, there are some scenarios where they take a decision even though they do not know the answer since the classification outcome only relies on the most activated neuron of the output layer. Ref. [23] demonstrated the need of evaluating the uncertainty of a model's predictions in order to improve the decisions of the system. Bayesian approaches offer a practical solution for understanding the uncertainty of the decisions of a neural network [24]. Specifically, they model a combination of aleatoric and epistemic uncertainty in order to increase loss robustness to noisy data, which usually leads to a boost in performance [23].

Despite the large performance provided by recent techniques, they do not usually provide a crucial information regarding an individual prediction is reliable or not. This is particularly important in contexts where taking a wrong decision can have a dramatic effect. To address this issue, in this work we employ an ensemble classification system based on a Bayesian Deep Learning approach in order to maximize performance while quantifying the uncertainty of each classification decision. In particular, we combine a number of Convolutional Neural Networks (CNNs) with the same structure, but differing in the kernel size of their convolutional layers. This allows the classification system to extract relevant features of different size and shape. The global classification is performed by combining the predictions of the different classifiers. The contribution of each individual classifier depends on the uncertainty of their predictions: the lower the uncertainty, the higher the weight, and vice versa. Performance of our proposal is evaluated in a range of real scenarios: from the analysis of two-dimensional images in order to distinguish between different types of pneumonia to the application of the proposed method to three-dimensional images in search of informative patterns associated with Parkinson's disease.

This work skillfully uses Deep Learning within a Bayesian framework for image classification, obtaining a large performance while quantifying the reliability of the model's predictions. The rest of the paper is organized as follows. Section 2 summarizes related works for image classification. Section 3 explains the different stages of the proposed method. First, the CNNs employed are presented. Then, the Bayesian framework used in combination with the neural networks is detailed. Finally, the multi-level ensemble approach for the multi-class classification is described. Afterwards, in Section 4, the proposed method is applied to chest X-ray (CXR) images in order to evaluate its suitability to the diagnosis of different types of pneumonia, in addition to be applied to SPECT images to assess its ability for the diagnosis of Parkinson's disease. Specifically, a description of the datasets and the preprocessing applied is included, as well as an explanation of the main experiments carried out in this work. Results are summarized in Section 7, whereas Section 8 includes a discussion of the findings. Finally, the conclusions and future works are available in Section 9.

2. Related works

The use of automatic systems for image classification is widespread. With reference to deep learning, a high number of methods have been proposed for this purpose. The emergence of AlexNet [25] was a shift of paradigm given the excellent performance achieved in the ImageNet LSVRC-2012 contest. Based on this architecture, [26] reduced the size of the convolutional kernels (and therefore, the complexity of the model) while obtaining better results than AlexNet. The appearance of inception modules [27] allowed to preserve a high performance while considerably decreasing the complexity of the networks. This is a crucial point because an increase in the depth of models can lead to a worse performance given the difficulties in optimization [28]. To mitigate this issue, [29] proposed the residual learning strategy (also known as ResNet) in order to ease the training of networks with a substantial complexity. Briefly, they introduced the concept of identity skip-connections. Unlike the traditional sequential pipeline, this allows

data to pass from any layer to any subsequent one [30]. The main advantage of adding this skip-connection is that any layer can be skipped if degrades the performance of the model. The combination of inception modules and residual blocks outperformed other models in the 2012 ImageNet challenge, in addition to highly accelerate the training of the model [31]. Following the development of this architecture, [32] added to Inception model a separable convolution component to perform convolutions in a more efficient way.

Explainable models emerged with the aim of providing interpretability to classification decisions. Intelligent systems are usually considered as black boxes with a lack of transparency that do not offer an explanation apart from the decision itself. This is specially relevant in deep learning approaches, where disentangling the relationships between inputs and outputs is not a straightforward task. There is a high number of approaches trying to provide explanations to deep neural networks [33]. Most of them are known as attribution-based methods, and they try to localize regions that contribute most to the decision giving explanation in the domain of the input images. One example is the Class Activation Mapping (CAM), an alternative which generates an attribution map for visually explaining the decision of an image classification [34]. Based on this method, [35] developed a generalization of CAM visualization that used gradients to generate class-discriminative visualization. In [36], authors developed an explainable skin cancer diagnostic system based on the taxonomic organization of skin lesions. This system identified different pathologies and provided the most probable spatial attention maps for each decision of the classification. Other methods have based on producing alterations to the inputs and comparing the outputs before and after these alterations to identify meaningful predictors [37,38].

Previous works have employed uncertainty as an informative measure about a model prediction. Specifically, uncertainty can denote the ambiguity and inaccuracy of a classifier decision when a model does not know the answer. Bayesian Deep Learning makes use of uncertainty to provide interpretable solutions. The applicability of this approach to medical images has been widely demonstrated in literature. In [39], a Bayesian approach was presented for the glaucoma diagnosis. This method combines information from multiple medical data sources and estimates the reliability of the classification prediction. Ref. [40] proposed a supervised Bayesian learning algorithm for cancer detection using terahertz imaging. Specifically, they used a reliability-based training data selection method in order to only use for training those data that exceed a certain threshold. Similar architectures have been used for coronary roadmapping in percutaneous coronary intervention [41], in the segmentation of diabetic retinopathy lesions [42] and in the segmentation of masses associated with breast cancer [43].

3. Methodology

3.1. Convolutional neural networks

The main aspect when developing an image classifier is the performance that the approach can lead to. As explained in previous sections, the emergence of algorithms based on deep learning revolutionized the automatic classification of images because of the high accuracy that they usually provide [44–47]. In medical images, CNNs have been widely used from its development for the ImageNet classification benchmark [48]. This architecture emerged as an attempt of replicating the behavior of neurons. Briefly, CNNs combine different steps based on convolution and pooling to allow the identification of different patterns and low and high-level features [44,49]. The main component of a CNN is known as convolutional layer. This operator takes the tensor $\mathbf{V}_{i-1}$ containing the activation map of the previous layer $i-1$. Thus, the target layer (i) learns a set of N filters $\mathbf{W}_i$ with a bias term $\mathbf{b}_i$, as follows:

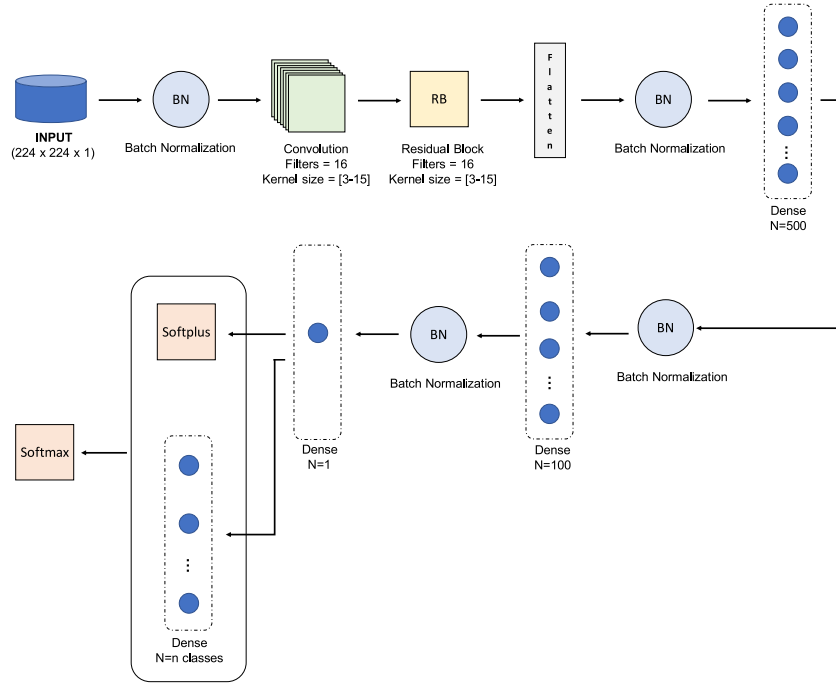$$\mathbf{V}_i = f_a(\mathbf{W}_i * \mathbf{V}_{i-1} + \mathbf{b}_i) \tag{1}$$

Fig. 1. Diagram of the bayesian framework of each individual network within the ensemble.

where $f_a(*)$ is the activation function [44]. For a two-dimensional environment ($\mathbf{V}_{i-1}$) of size $H \times W \times C$ (height, width and number of channels, respectively), $\mathbf{W}_i$ is of size $P \times Q \times S \times K$ where K is the number of filters. The kth convolution term for the kth filter is

$$\mathbf{W}_{ik} * \mathbf{V}_{i-1} = \sum_{u=0}^{P-1} \sum_{v=0}^{Q-1} [\mathbf{W}_{ik}(P-u, Q-v) \qquad (2)$$
$$. \mathbf{V}_{i-1}(x+u, y+v)]$$

Once convolution is performed, the activation of the filters in layer i are stored and passed to the next layer $i + 1$. We employed a version of the ResNet-18 presented in [50] as the base CNN. The output layer contained 2 neurons with softmax activation. Besides, dropout was used to prevent overfitting, and Batch Normalization for convergence. The number of filters in each layer is an important parameter to be set: the higher this number, the more patterns the model is able to learn. There is no consensus in literature about the ideal number of filters, probably because different problems need CNNs with different configurations, but numbers that are a power of 2 are usually taken. We employed 16 filters in each convolutional block in order to strike a balance between performance and computational burden. Fig. 1 shows a schematic representation of the CNN proposed.

3.2. Bayesian deep learning

Most of the classification studies evaluate their performance in terms of accuracy, area under the ROC curve or other measures derived from the confusion matrix. When modeling two classes such as in binary classifications, a test sample is always labeled as belonging to one of the two classes. This means that although the test sample does not belong to any of these classes, it would be assigned to one of them. This situation is not problematic because the development of a classification system takes into account the classes to be modeled. In most of scenarios, it makes no sense to introduce a test sample from a different class to the ones used for fitting the model. However, it would be highly useful to evaluate the reliability of a model's prediction in order to identify situations where the classifier does not know the answer. Recent works have claimed the need of computing the uncertainty of a prediction, allowing its rejection when the uncertainty is too high. To address this issue, Bayesian modeling can be used for understanding uncertainty associated with deep learning models [51]. According to the Bayesian framework, there are two types of uncertainty that can be estimated: epistemic and aleatoric [24,52]. Epistemic uncertainty is inherent to the model, which means that it can be reduced by increasing the processed data. Estimating this kind of uncertainty requires to model distributions over the different parameters of the model. This allows to optimize the network according to the average of all possible weights. Let $\mathbf{x}$ be a feature vector and $\mathbf{W}$ the weights of a Bayesian Neural Network (BNN). Considering the output of the network as $\mathbf{f}^{\mathbf{W}}(\mathbf{x})$, the model likelihood can be defined as $p(\mathbf{y}|\mathbf{f}^{\mathbf{W}}(\mathbf{x}))$. For a given dataset $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, the Bayesian inference computes the posterior probability over the weights $p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$.

Aleatoric uncertainty is usually referred as the uncertainty inherent to the data, and can be divided into two sub-categories: (i) homoscedatic uncertainty, which remains stable for every input of the model; and (ii) heteroscedastic, which assumes that noise varies for the different inputs of the model [53,54]. Heteroscedastic uncertainty can be modeled by modifying the loss function used by the neural network. Since this uncertainty is a function of the input data, employing a deterministic mapping from inputs to model outputs can allow the estimation of the uncertainty. For a typical Euclidean loss $L = \|y - \hat{y}\|^2$, the Bayesian version will be given by $L = \frac{\|y-\hat{y}\|^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2$. In the latter one, the model predicts both $\hat{y}$ and variance σ^2, so that if model prediction is not good, the residual term will be attenuated by increasing σ^2. Therefore, the term $\log\sigma^2$ prevents uncertainty growing until infinite, leading to a learned loss attenuation. The process for homoscedastic uncertainty is essentially the same, but considering the uncertainty like a free parameter instead of a model output.

We modified the ResNet-18 described in Section 3.1 to obtain a Bayesian version of this CNN. Specifically, we replaced the deterministic weights along the network by a distribution over these parameters. Ref. [55] demonstrated that applying dropout before every weight layer in a neural network is mathematically equivalent to an approximation to the probabilistic deep Gaussian process [56]. Briefly, they showed that the dropout objective minimizes the Kullback–Leibler divergence between an approximate distribution and the posterior one of a deep Gaussian process. We applied a Monte Carlo dropout sampling to

place a Bernoulli distribution over the network's weights during the test phase. This allows to obtain a distribution for the output predictions [57,58]. The statistics of this distribution reflect the model's uncertainty. As a result, the loss function depends on two factors: the softmax values (as in the non-Bayesian modality) and the Bayesian categorical cross entropy, which is based on the input variance (see [23] for more details).

3.3. Multi-level ensemble classification

There are some situations in which combining several classifiers provide more accurate results than employing an individual one. This is known as ensemble classification, and it is particularly useful for fusing data from different modalities or when different architectures are simultaneously employed [7,8]. Previous studies have used majority voting to combine the output of the base classifiers [59,60]. Given the Bayesian nature of the approach presented in this work, we propose to employ the uncertainty of the prediction of each individual network to weight their contribution to the ensemble. This means that if the uncertainty of a classifier in a specific prediction is high, the contribution to the final ensemble will be low, and vice versa [61]. Defining $u_l^k(\mathbf{y})$ as the uncertainty of the test sample $\mathbf{y}$ obtained from the kth classifier corresponding to the lth class, the empirical average of the lth weights (inverse of uncertainties) over the K classifiers can be calculated as follows:

$$E_l(\mathbf{y}) = \frac{\sum_{k=1}^{K} \frac{1}{u_l^k(\mathbf{y})}}{K} \tag{3}$$

The class label of the test sample $\mathbf{y}$ is then assigned to the class with the maximum average weight as:

$$Label(\mathbf{y}) = \underset{l}{\mathrm{argmax}}\ E_l(\mathbf{y}) \tag{4}$$

We employed this approach for weighting the contribution of each member of the ensemble in order to guarantee that the classification decision is guided by the reliability of the decision, and not derived only from performance, as other much common ensemble frameworks use. Identifying the patterns associated with two classes is an interesting initial step in the development of a CAD system. However, there are some contexts in which it is much more useful to differentiate between a number of classes. In order to perform the multiclass classification, we employed a decision tree based on the One-versus-all (OVA) approach [62–64]. This alternative divides a multiclass problem into a number of binary sub-problems. In each one of them, one of the classes is considered as the positive class, whereas the other classes are the negative class. Thus, the proposed method in this work provides a solution that informs about the reliability of a prediction. Besides, the system uses this uncertainty to assign a weight to each member of the ensemble, in addition to be compatible with multiclass classification.

Fig. 2 shows a schematic representation of the entire classification framework. The CNNs differ in the kernel size of their convolutional layers and are combined within an ensemble method. The uncertainty in the predictions of each individual classifier are then used to weigh their contribution to the final decision. This process can be summarized as follows:

- Definition of a CNN that adapts to the idiosyncrasy of the specific classification problem. We selected the ResNet-18 as the base CNN since its performance has been widely evaluated in literature.
- Modification of the CNN in order to obtain a Bayesian version. This is done by replacing the deterministic weights along the network by a distribution over these parameters within a Monte-Carlo dropout sampling. Thus, the CNN provides two different outputs: the classification decision, and the uncertainty of this decision.
- The decision of each CNN is combined within an ensemble framework. The contribution of each individual CNN to the final decision is given by the reliability of their predictions.

4. Application to pneumonia detection

In this section, we evaluate the applicability of our proposal as a tool for the automatic classification of chest X-ray images. Specifically, we assess the performance of our method in order to identify the presence of pneumonia, as well as distinguishing whether it is caused by viral or bacterial pathogens.

4.1. Database description

We have used the dataset available in [65] for controls and patients who suffered from a bacterial or a viral pneumonia. According to the information described in [66], the CXR images were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All CXR images were obtained as part of patient's routines clinical care. Institutional Review Board (IRB)/Ethics Committee approvals were obtained. The work was conducted in a manner compliant with the United States Health Insurance Portability and Accountability Act (HIPAA) and was adherent to the tenets of the Declaration of Helsinki. [66] collected and labeled a total of 6374 CXR images from children, including 4273 characterized as depicting pneumonia and 1583 normal. From those patients diagnosed with pneumonia, 2786 were labeled as bacterial pneumonia, whereas 1487 were labeled as viral pneumonia. Fig. 3 shows the CXR image from a control (CTRL), and a patient suffering from a bacterial (BAC) and a viral (VIR) pneumonia (see Fig. 3).

4.2. Image preprocessing

When working with medical images, it is crucial to apply a preprocessing that improves the subsequent classification performance. This is especially important in CXR images, where low X-ray radiation and movement during image acquisition result in noisy and low-resolution images. However, this preprocessing must adapt images to the needs of the neural network. Due to computational and memory requirements, we downsampled the input images to obtain a final map of size 224×224. We also performed an intensity normalization procedure for each individual image based on standardization. Each image was transformed such the resulting distribution has a mean (μ) of 0 and a standard deviation (σ) of 1, as follows:

$$I' = \frac{I - \mu}{\sigma} \tag{5}$$

where I is the original image and I' is the resulting one.

4.3. Ensemble of Bayesian CNNs

As explained in previous sections, we propose the use of a method based on an ensemble of Bayesian CNNs to obtain a measure of the uncertainty of the predictions. The election and setup of the architecture was done according to the database to be analyzed. In our case, the images correspond to CXR images from patients with different lung pathologies. Patterns associated with each type of pneumonia are similar among different subjects. However, there are some factors like the virulence of the disease and the presence of other pulmonary findings that can affect the identification of the patterns associated with the different pathologies. To overcome this issue, we employed an ensemble of seven neural networks, each one of them with a different kernel size value in the range $[3-15]$ with increments of two. This means that the kernel size assigned to the first network was 3, 5 for the second network and so on, until a size of 15 for the seventh CNN. The number of neural networks and their kernel sizes were selected in order to strike a balance between performance and computational cost.

Regarding the multiclass classification (controls *vs.* bacterial pneumonia *vs.* viral pneumonia), we used a decision tree with two levels in order to distinguish between the different pathologies. In each level, an ensemble of different kernel sizes was employed. This led to a two-level
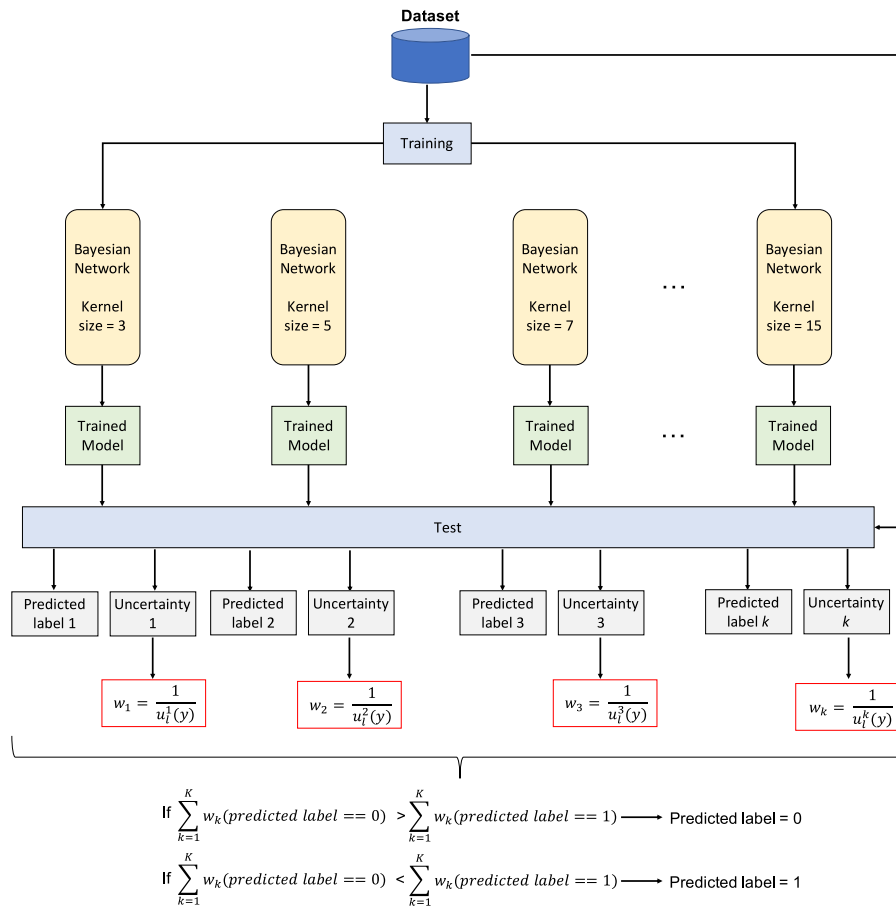
Fig. 2. Schema of the ensemble architecture proposed in this work based on the uncertainty in the prediction of each individual classifier.
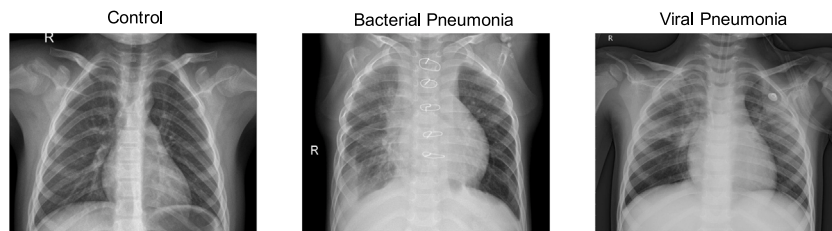

Fig. 3. Representation of the different pathologies evaluated. From left to right, CXR image of a control, bacterial and viral pneumonia.

ensemble classification: one ensemble for the combination of different kernels, and another one for combining binary classifiers to perform multiclass classification. The decision tree relies on a process that can be summarized as follows:

- First level: classification between normal *vs.* pneumonia. The second class contains subjects diagnosed from the two different types of pneumonia (bacterial and viral).
- Second level: classification between bacterial *vs.* viral pneumonia.

Fig. 4 depicts a visual representation of how the decision tree works. Images that are labeled as pneumonia in the first level are passed to the second one. Here, it is identified whether the virus that produced the pneumonia was a virus or a bacteria. It is worth mentioning that the binary classifier employed in each level has the same ensemble structure that the one explained in Section 3.3.

5. Application to Parkinson's diagnosis

We also evaluate the applicability of our method as a diagnostic tool of Parkinson's disease (PD) from SPECT (Single Photon Emission Tomography) images. Specifically, we assess the ability of our proposal to differentiate between the striatal region of PD patients and healthy individuals.

5.1. Database description

The data used in the preparation of this article was obtained from the PPMI (Parkinson's Progression Markers Initiative, RRID: SCR_006431). PPMI is an observational clinical study to verify progression markers in PD. For up-to-date information on the study, visit https://www.ppmi-info.org/. Raw projection data are acquired into a 128×128 matrix stepping each 3 degrees for a total of 120 projection into two 20% symmetric photopeak windows centered on 159 keV and 122 keV with a total scan duration of approximately 30–45 min [67]. 269 DaTSCAN images have been used in this study, 158 from patients suffering PD disease and 111 normal controls (CTRL). Fig. 5 provides a visual example of the DaTSCAN images of a control and a Parkinson's disease patient (left and right, respectively).
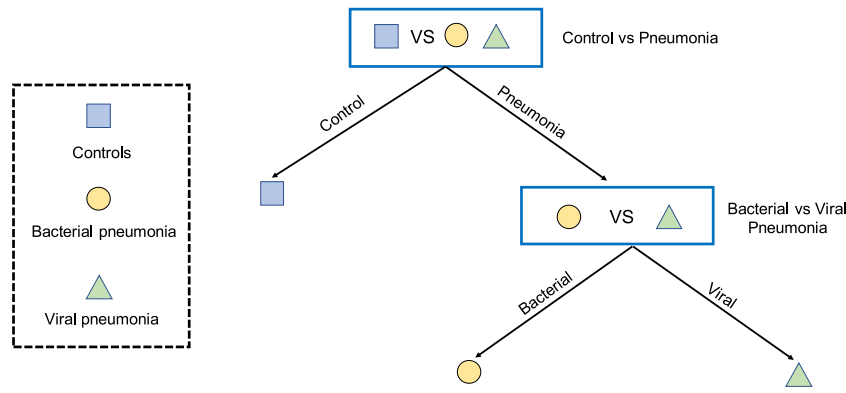
Fig. 4. Schematic representation of the decision tree employed for the multiclass classification.
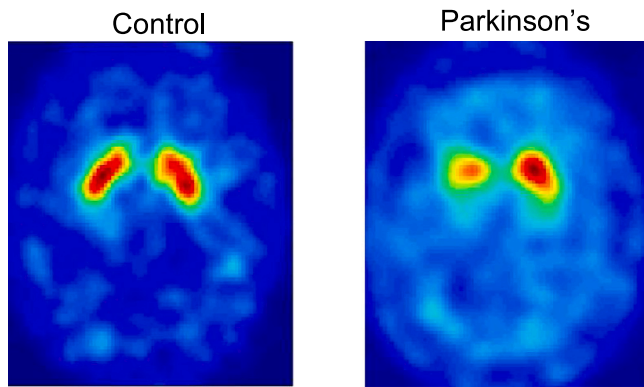

Fig. 5. Slice of the 3D images corresponding to a control (left) and a PD patient (right).

5.2. Spatial normalization

Spatial normalization is widely used in neuroimaging studies, and it is specially common when performing group analysis. The aim of this operation is to reduce the anatomical differences (i.e. shape and size) of the individual brains of the different patients. To do so, individual images are mapped from their original space to a new and common reference one, which is usually given by a template. We employed the most frequent one, proposed by the Montreal Neurological Institute, which is based on the average of 152 MRI scans from normal subjects. The original DaTSCN images extracted from the PPMI dataset preserved their original anatomical information. We used SPM12 to achieve the best warping of the images with the DaTSCAN template defined in [68]. After that, we isolated the regions of interests, which are given by those that reveal dopaminergic activity. This process leads to a high reduction in the size of the images (from (95,69,79) to (29,25,41)), which allows a large decrease in the computational cost while preserving the informative regions.

5.3. Intensity normalization

Another crucial aspect is related to the intensities of the preprocessed images. The idea behind SPECT images is that the intensity of the pixels provides an indirect measure of the neurophysiological activity. Thus, similar intensity values should correspond to similar drug uptakes, whereas differences in these values could reveal different pathologies [69–72]. In order to ensure that intensity levels are related to drug uptakes, we used an Integral Normalization [73], as follows:

$$\hat{\mathbf{I}}_i = \frac{\mathbf{I}_i}{I_{n,i}} \tag{6}$$

where $\mathbf{I}_i$ refers to the image of the ith subject in the database, $\hat{\mathbf{I}}_i$ is the resulting normalized image, and $I_{n,i}$ is the intensity normalization value. This is computed as the mean of the image for each independent subject. Finally, the resulting values were standardized in the range [0,1].

5.4. Ensemble of Bayesian CNNs

As explained in Section 4.3, the setup of the proposed method must adapt to the idiosyncrasy of the classification context. To do so, we employed an ensemble of five neural networks with different kernel sizes, from 2 to 6. It is important to note that SPECT are three-dimensional images, whereas CXR employed in pneumonia detection are two-dimensional. For this reason, we modified the convolutional and pooling layers of the CNN in order to provide a three-dimensional version of our Bayesian ensemble. It is worth mentioning that the additional dimension of these images highly increases the computational burden of the analysis. We used five CNNs instead of the seven employed in the classification of pneumonia in order to mitigate the computational cost associated with the processing of 3D images while preserving a high performance.

6. Performance evaluation

For all experiments, a 5-fold stratified cross-validation scheme was used to estimate the generalization ability of our method [74]. The performance of the classification frameworks was evaluated in terms of different parameters from the confusion matrix, which can be computed as follows:

$$Acc = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad Sens = \frac{T_P}{T_P + F_N}$$
$$Spec = \frac{T_N}{T_N + F_P} \quad Prec = \frac{T_P}{T_P + F_P} \tag{7}$$
$$F1-score = \frac{2 \times Prec \times Sens}{Prec + Sens}$$

where T_P is the number of pneumonia patients correctly classified (true positives), T_N is the number of control patients correctly classified (true negatives), F_P is the number of control subjects classified as pneumonia (false positives) and F_N is the number of pneumonia patients classified as controls (false negatives). We also employed the area under the curve ROC (AUC) as an additional measure of the classification performance [75,76]. Since classes were unbalanced (e.g. the number of pneumonia patients was higher than controls), we incorporated the weights of the classes into the cost function in order to the majority class does not contribute more than the minority one.

Given the ensemble nature of the system proposed in this work, we employed a kappa-uncertainty diagram to evaluate the level of agreement of the different classifier outputs while correcting for chance [77, 78]. This measure is based on Cohen's kappa coefficient [79], which is

widely accepted as the de facto standard for measurement of interannotator agreement [80]. Specifically, the kappa statistic compares an observed accuracy with an accuracy obtained by chance, providing a measure of how closely instances classified by a classifier match the ground truth. Mathematically, Cohen's kappa can be defined as:

$$k = \frac{p_A - p_E}{1 - p_E} \tag{8}$$

where p_A is the observed relative agreement between two annotators, and p_E is the probability of agreement by chance. Although acceptable kappa statistic values vary on the context, the closer to 1, the better the classification. Section 7 summarizes the kappa scores obtained by different members of the ensemble classifier, as well as revealing the relationship between the uncertainty of Bayesian networks and kappa values.

As explained in Section 3.3, a decision tree was employed for multiclass classification. In order to build the kappa-uncertainty diagram explained above, a combination of the uncertainties of the different levels of the tree has to be computed. To do so, we employed a method known as summation in quadrature [81], described as follows:

$$u_c(y) = \sqrt{\sum_{i=1}^{n}[c_i u(x_i)]^2} \tag{9}$$

where $u_c(y)$ is the combined uncertainty, c_i is the sensitivity coefficient and $u(x_i)$ is the standard uncertainty.

6.1. Experimental setup

6.1.1. Pneumonia detection

We define two experiments associated with the identification of pneumonia:

- **Experiment 1: Binary Classification** between different groups under two scenarios: **CTRL *vs.* PNEU**, which includes all images labeled as CTRL and PNEU; **BAC *vs.* VIR**, which divides the images from people diagnosed from pneumonia regarding the cause of the disease is a bacteria or a virus. The whole Bayesian CNN was trained using the Adam optimization algorithm [82], with learning rate 0.001, $\phi = 0.9$ and a decay of 0.001. The number of epochs employed for training the system was 15, 20 for the CTRL *vs.* PNEU and BAC *vs.* VIR, respectively.

- **Experiment 2: Multiclass Classification** by using a decision tree in order to distinguish between the three different pathologies contained in the database. A binary classification is employed in each of the two levels of the tree. The first level corresponds to the CTRL *vs.* PNEU classification, whereas the second one contains the BAC *vs.* VIR comparison. These binary classifiers employ the same framework and configuration as in Experiment 1.

Table 1 provides an overview of recent work focused on the automatic detection of pneumonia.

6.1.2. Parkinson's diagnosis

Experiment 1: Classification between PD patients and normal controls. The Bayesian ensemble of five CNNs was trained using the Adam optimization algorithm [82], with learning rate 0.001, $\phi = 0.9$ and a decay of 0.001. Besides, the system was trained during 30 epochs. Table 2 summarizes the results obtained by recent works in the automatic diagnosis of Parkinson's disease.

The experiments conducted in both scenarios (pneumonia detection and Parkinson's diagnosis) employ custom code. This was written in Python 3.6, and a number of additional libraries was used: Keras library over Tensorflow and Numpy 1.19.5. The experiments were carried out on a cluster with the following hardware specifications: two Intel® Xeon® E5-2630 node 2.40 GHz processors, with 10 cores per processor. Besides, the total RAM memory capacity of the system is 128 GB.

Regarding the execution time, it is quite dependent on the classification scenario, whereas parameters such as the kernel size also affects the computational cost. On average, our classification framework requires 5 h and 20 min for the pneumonia detection context, and 6 h and 15 min for the Parkinson's diagnosis. Although the number of images in this last context was lower than in the first one, the processing of 3D images considerably increased the computational load, balancing the time needed for performing the analyses.

7. Results

7.1. Pneumonia detection

We first explore how performance varies for the different kernel sizes of the individual classifiers for all the binary classifications performed (see Fig. 6). We can see that kappa score slightly varies when increasing the kernel size in the two classification contexts. With reference to uncertainty, only in the BAC *vs.* VIR scenario uncertainty values drastically change for different kernel sizes. Therefore, there is not a tendency that let us assure that there is a relationship between these two variables. It is important to note the high levels of uncertainty in this classification context when comparing to the first one, which manifests the extreme difficulty of this specific classification. It is not surprising that differentiating between a control and a patient who suffer from pneumonia is a considerably easier task.

We observe that the discrimination ability of the system is very high for the two binary classifications regardless of the kernel size employed. Results in terms of different performance measures are shown in Table 3, whereas Fig. 7 depicts the ROC curves for the different classifiers. Large values are obtained, as expected, in the CTRL *vs.* PNEU context. However, these results confirm that our system can also separate patients with the same diagnosis (pneumonia) but with a different cause (bacteria, virus). We also use the kappa-uncertainty diagram to evaluate the level of agreement between the classifier outputs. Fig. 8 shows these diagrams for the two binary classifiers and the multiclass derived from the decision tree, represented by a different color. The cloud points represent the kappa score-uncertainty obtained in each fold of the cross-validation scheme, whereas large stars represent the centroid of the resulting distribution. From this figure, we can see that there is not a great difference between individual classifiers, in consonance with results derived from ROC curves.

It is interesting how this figure reveals that the combination of classifiers with a certain performance (high kappa score and low uncertainty) leads to an ensemble classifier with these features. However, uncertainty is higher in the multiclass classifier for a similar kappa score compared to individual ones. This means that, although classification performance of the decision tree is high, the uncertainty of the resulting prediction is also higher than in binary classification. This evidences the extreme utility of this kind of diagrams in Bayesian deep learning and in contexts when reliability of predictions is of core interest. According to Table 3, the multiclass classifier has a superior performance than the CTRL *vs.* PNEU in some of the evaluated metrics. However, the uncertainty of the predictions is also higher (centroid of the multiclass is farther to the right than the CTRL *vs.* PNEU centroid). Further discussion regarding the results obtained and their clinical implications are provided in Section 8.

7.2. Parkinson's diagnosis

Fig. 6(c) shows the relationship between performance and the kernel sizes when trying to distinguish controls and PD patients. We observe that the kappa score remains stable regardless the size of the kernels of each individual CNN. Besides, only the kernels with the largest sizes (5 and 6 voxels) provide a less reliable prediction according to the increase in the uncertainty obtained. Given the small size of the images it is not uncommon that large kernels lead to a poorer

Table 1

Summary of previous works focused on the automatic identification of pneumonia.

Research work	Dataset	Method	Classification context	Results (%)
[83]	852 CXR images	COVIDNet	Normal *vs.* COVID	Acc = 97.72
[84]	521 CT images	Pretrained model	Normal *vs.* COVID *vs.* Tuberculosis	Acc = 97.32
[85]	161 Lung ultrasound	CNN	Normal *vs.* COVID *vs.* Bacterial pneumonia	Acc = 91.80
[86]	1110 CT scans	COV-CAF	Normal *vs.* COVID	Acc = 97.76
[87]	640 CT images	FGCNet	Normal *vs.* COVID	Acc = 97.71
[88]	420 CXR images	Machine learning	COVID *vs.* Non-COVID pneumonia	Acc = 94.00
[89]	194 922 CCT images	Transfer learning	Normal *vs.* COVID *vs.* Non-COVID pneumonia	Acc = 99.00
[90]	1142 CXR images	DarkNet	Normal *vs.* COVID *vs.* Viral pneumonia	Acc = 87.02
[91]	3487 CXR images	DenseNet-201	Normal *vs.* Bacterial *vs.* Viral pneumonia	Acc = 97.94
[92]	6214 CXR images	CoVNet-19	Normal *vs.* COVID *vs.* Pneumonia	Acc = 98.33
[93]	2905 CXR images	InstaCovNet-19	Normal *vs.* COVID *vs.* Pneumonia	Acc = 99.08
[94]	3522 CXR images	CHP-Net	Normal *vs.* COVID *vs.* Pneumonia	Acc = 93.65
[95]	16 634 CXR images	EfficientNet	Normal *vs.* COVID *vs.* Other diseases	Acc = 93.48
[96]	1218 CXR images	DenseNet-201	COVID *vs.* Pneumonia	Acc = 94.96
[97]	230 CXR images	ConvNet	Normal *vs.* COVID *vs.* Viral pneumonia *vs.* Bacterial pneumonia	Acc = 95.60
[98]	13 975 CXR images	CoV2-Detect-Net	Normal *vs.* COVID *vs.* Pneumonia	Acc = 99.34
[99]	112 120 CXR images	CheXNet	14 diseases	Acc = 87.00
[100]	1125 CXR images	SLB and FFB Nets	Normal *vs.* COVID	Acc = 99.52
[101]	1508 CXR images	EfficientNet	Normal *vs.* COVID *vs.* Pneumonia	Acc = 99.62

Table 2

Summary of previous works focused on the automatic identification of pneumonia.

Research work	Dataset	Method	Classification context	Results (%)
[102]	80 sMRI	SVM-RBF	PD *vs.* CTRL	Acc = 87.5
[103]	182 PET	SVM-sigmoid	PD *vs.* PD *vs.* CTRL	Acc = 91.26
[104]	624 SPECT	CNN	PD *vs.* PD *vs.* CTRL	Acc = 96.00
[105]	642 SPECT	CNN	PD *vs.* CTRL	Acc = 95.20
[106]	101 T1W1/DTI	Ensemble	PD *vs.* CTRL	Acc = 96.88
[107]	153 TCS	DNMLDM	PD *vs.* CTRL	Acc = 81.70
[108]	379 SPECT	CNN	PD *vs.* CTRL	Acc = 86.00
[44]	642 SPECT	CNN	PD *vs.* CTRL	Acc = 94.10
[45]	269 SPECT	CNN	PD *vs.* CTRL	Acc = 95.10
[109]	98 PET/CT	CNN	PD *vs.* CTRL	Acc = 93.00
[110]	406 MRI	CNN	PD *vs.* CTRL	Acc = 95.30
[111]	341 PET	CNN	PD *vs.* CTRL	Acc = 84.20
[112]	200 SPECT	CNN	PD *vs.* CTRL	Acc = 85.00
[113]	230 MRI	CNN	PD *vs.* CTRL	Acc = 81.00

Table 3

Performance of the ensemble classification approach proposed in this work in the different contexts evaluated. In the Parkinson's diagnosis scenario, we employed as baseline voxels as features (VAF) and a Linear Support Vector Machines (SVM). The optimum parameters were computed within a grid-search procedure.

| Experiment	Acc (%)	Sens (%)	Spec (%)	Prec (%)	AUC (%)	F1-score (%)
Pneumonia detection						
CTRL vs. PNEU	97.21 ± 2.87	96.63 ± 4.28	99.77 ± 0.18	99.92 ± 0.09	98.31 ± 2.15	98.22 ± 2.55
BAC vs. VIR	98.29 ± 0.91	98.25 ± 1.11	98.34 ± 0.67	98.77 ± 0.76	98.55 ± 0.98	98.37 ± 0.59
Multiclass	98.19 ± 1.50	97.18 ± 2.76	99.25 ± 0.32	99.55 ± 0.28	98.45 ± 1.41	98.54 ± 1.35
Parkinson's diagnosis						
CTRL vs. PD (Baseline VAF+SVM)	90.63 ± 1.23	91.23 ± 0.87	89.96 ± 1.76	90.25 ± 1.21	91.32 ± 1.43	90.57 ± 1.39
CTRL vs. PD (Our method)	95.31 ± 1.75	94.36 ± 1.34	95.76 ± 1.27	94.83 ± 1.62	95.55 ± 1.18	94.87 ± 1.41

performance. This is one of the reasons why we used relatively small kernel sizes, in addition to the huge computational cost of convolution in three-dimensional images. However, even in these extreme scenarios the uncertainty is lower than in the context of pneumonia diagnosis, evidencing the good performance of our method when processing this kind of images. This ability for identifying the presence of Parkinson's disease can also be shown in Table 3, leading to an accuracy of 95.31%. Moreover, the ROC curve depicted in Fig. 7(b) provides a visual evidence that the proposed system can be successfully applied in this clinical context.

8. Discussion

In this study, a classification framework based on Bayesian Deep Learning is proposed. This approach is based on the quantification of the uncertainty of the model's prediction. Specifically, we employed an ensemble scheme in order to combine several Convolutional Neural Networks with the same structure but different kernel size in the convolutional layers. This allows extracting information from features of different size and shape. The decision of each individual CNN is fused according to the reliability in the prediction of each one of them, so that a prediction with a low uncertainty has a higher weight in the final decision than a prediction with a high uncertainty. This alternative was applied to two different datasets in order to evaluate its performance in the classification of two-dimensional and three-dimensional images. In the first dataset (pneumonia detection), three scenarios with differential difficulty were evaluated: (i) the two classes generated relatively big differences in the observed pattern (control *vs.* pneumonia); (ii) the differences between the classes were lower (bacterial *vs.* viral pneumonia); (iii) a multiclass context where simultaneously differentiating between three different pathologies. In the second dataset, the aim was to differentiate between the striatal region of PD patients and healthy individuals.
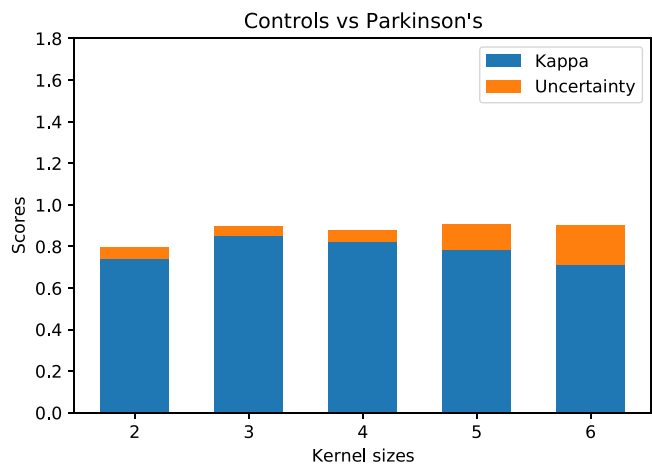
A high number of studies have employed Deep Learning for image classification because of the high performance that usually provides. However, most of these studies only evaluate the performance itself,
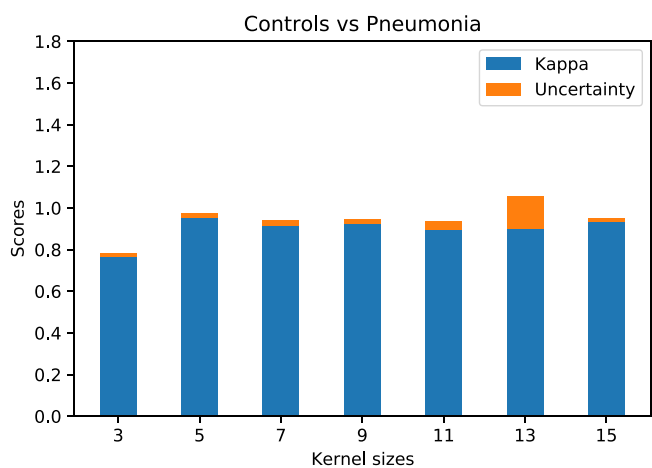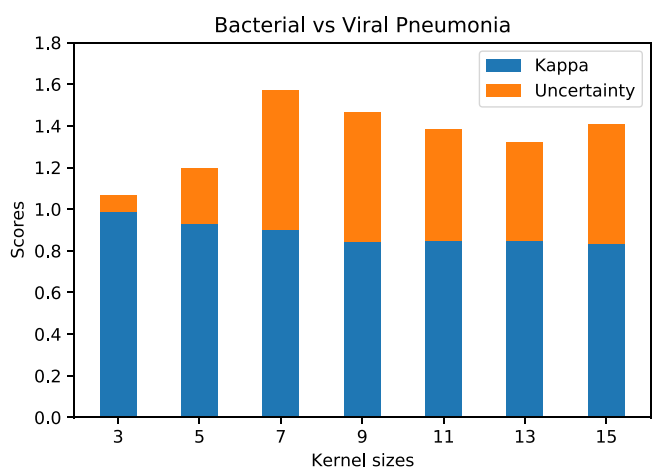
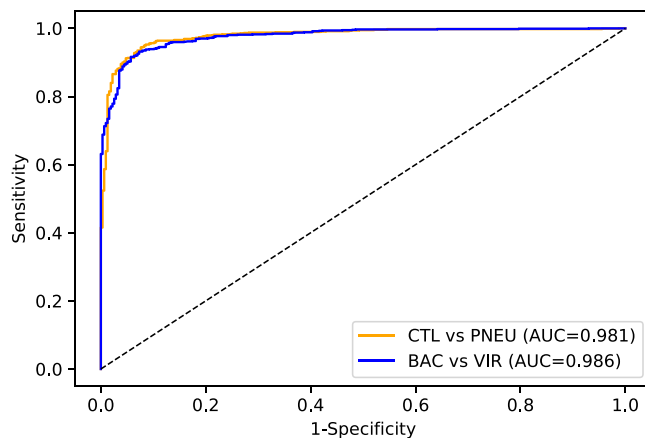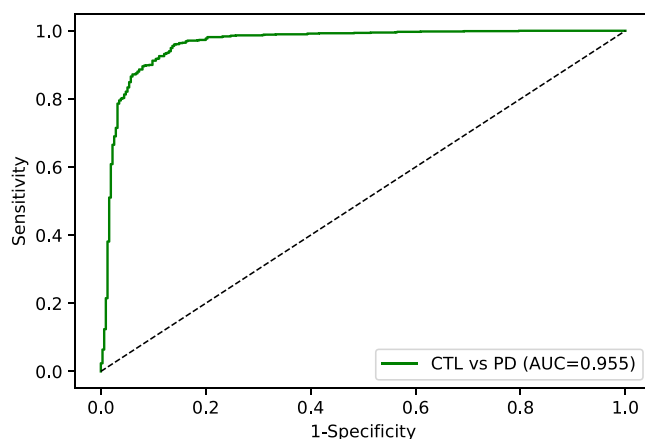**Fig. 6.** Performance associated with the different kernel sizes in terms of kappa and uncertainty for the different classification scenarios: (a) Controls *vs.* Pneumonia; (b) Bacterial *vs.* Viral pneumonia; (c) Controls *vs.* Parkinson's disease.

Fig. 7. ROC curves obtained in the different classification scenarios: (a) Pneumonia detection; (b) Parkinson's diagnosis.
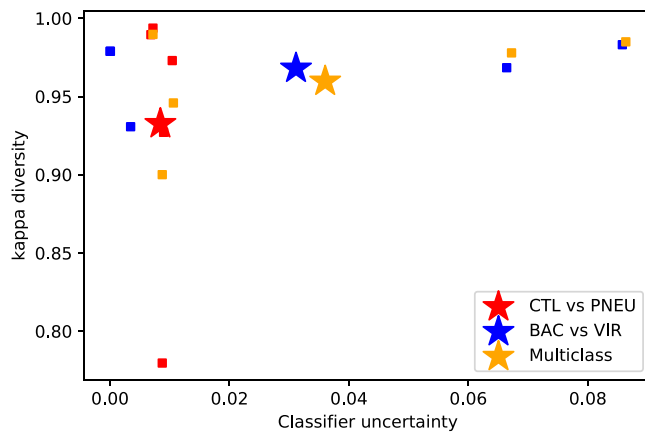

Fig. 8. Diversity-uncertainty diagrams of the different levels of the multiclass classifier. The *x*-axis represents the combined uncertainty of each individual classifier and the resulting multiclass. The *y*-axis represents diversity of the classifiers evaluated by the kappa measure. Each dot represents the kappa-uncertainty score obtained by a classifier in one fold, whereas large stars represent the centroid of the resulting distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and not how reliable the prediction is. The Bayesian framework allows the quantification of the uncertainty of a prediction, enhancing the usefulness of a CAD system in real scenarios. Most important, the information that uncertainty provides is even more useful when applied in combination with ensemble classification. In fact, computing the weight of each individual classifier according to the reliability of its predictions offers an alternative to other fusing methods that are more computationally demanding (those that are based on the accuracy computed from an inner CV loop).

The high performance shown by the proposed method in all scenarios manifest its large suitability for medical image classification, both for two-dimensional and three-dimensional images. With reference to 2D images, our system is able to detect the presence of pneumonia in CXR images, in addition to distinguish whether the source of the pathology is viral or bacterial. The features extracted by convolutional blocks of different kernel sizes contained relevant information that enhanced the separability between the different classes. The combination of this information is especially interesting in this context where the database contains images from a different virulence of the disease. Pulmonary affections caused by the different pathologies evaluated in this work mainly depend on the severity of the disease. The ensemble method proposed in this work allows the identification of patterns associated with pneumonia without focusing on a specific size for the informative regions. This crucial characteristic of our system is also relevant for the PD diagnosis, since it is possible to detect the pathology without inferring a specific magnitude of the striatal activity.

Another crucial aspect of the method proposed in this work is its Bayesian nature. The aim of CAD systems regardless of the application context is to maximize the classification performance, in terms of accuracy, AUC, etc. However, in most scenarios it is also important to know the reliability of the prediction itself. Neural networks are prone to overfitting, which means that taking decisions based only on the prediction can be counterproductive. In an extreme case, it is possible that the classifier does not know the class a test image belongs to, but it always has to assign a label, even though the output probability is near to chance level. This is particularly problematic when developing a tool for the diagnosis of a disease. Doctors need to know not only the global accuracy obtained during the training and test of the model, but how reliable the prediction of new individual samples is. This problem is addressed with the inclusion of Bayesian elements in neural networks. However, our findings reveal that this is not the only advantage that this approach provides. We have demonstrated the high performance of ensemble classification, even in situations where differences between the patterns associated with the different pathologies are extremely small. The novelty of our approach relies on the way the contribution of each individual classifier to the global decision is computed. Weights are usually derived from the accuracy of each individual classifier. However, results can be biased if part of the predictions are obtained by chance, i.e. when the output probabilities of the different classes are almost equal. We overcome this problem by weighting the contribution of each classifier according to the uncertainty of its predictions.

We have developed a tool that is able to distinguish between patterns associated with different pathologies, but it is worth highlighting the high performance obtained in the multiclass classification of the pneumonia detection. In this case, the accuracy and the AUC obtained were 98.19% and 98.45%, respectively, which is considerably higher than the results provided by similar techniques in previous studies [10, 60,114–116]. There are two main relevant aspects derived from these excellent results to be mentioned. First, the only preprocessing applied to the data was the rescaling of the images to a lower resolution in order to reduce the computational burden of the classification pipeline. We did not perform other complex processes such as lung segmentation, but the RAW rescaled images were used as the inputs of the classification system. Thus, it is remarkable the high performance obtained by the method proposed in this extreme situation. Second, results obtained in the multiclass scenario allow the application of the tool proposed

in this work in a more real context. The multiclass scenario is more similar to a real context than binary classifications, where the simplest case only distinguishes the presence (or not) of pneumonia. Results obtained by the multiclass classifier reveal the usefulness of our method. Regarding the results obtained in the diagnose of PD, the accuracy and AUC obtained were 95.31% and 95.55%, respectively, which is higher than most methods presented in recent works [103,105,108,111,112]. With reference to the preprocessing of SPECT images in PD diagnosis, the main difference is that images were coregistered, unlike CXR images. This is the main reason why the baseline alternative evaluated shows reasonable results. However, our method highly overcomes these results, evidencing that it can also be successfully applied to 3D images.

9. Conclusion

In this paper, we propose an ensemble method for multiclass classification based on uncertainty. This is addressed by employing a number of CNNs within an ensemble framework, where each CNN has the same architecture but differs in the size of the kernels in the convolutional layers. After training each network, the decision of each individual one are combined according to the reliability of their prediction: the lower the uncertainty, the higher the contribution, and vice versa. Finally, a decision tree is used by dividing the multiclass problem into a number of binary sub-problems. We obtained a 98.19% of accuracy and AUC of 98.45% in the multiclass scenario, when different types of pneumonia are distinguished, whereas Parkinson's disease was successfully diagnosed with a 95.31% of accuracy and AUC of 95.55%. The combination of CNNs of different kernel sizes allows the identification of informative patterns regardless of their size and shape. Moreover, the reduced preprocessing needed for obtaining these results guarantees a limited computational cost. Our findings validate the usefulness of the proposed method in a medical context, both for two-dimensional and volumetric images. In fact, the additional information that our method provides (obtained from the uncertainty of the predictions) paves the way to employ this methodology in a wide range of scenarios, especially where prediction itself is not enough to take a reliable decision.

CRediT authorship contribution statement

Juan E. Arco: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Andrés Ortiz:** Conceptualization, Methodology, Investigation, Writing – original draft. **Javier Ramírez:** Conceptualization, Methodology, Investigation, Writing – original draft. **Francisco J. Martínez-Murcia:** Conceptualization, Methodology, Investigation, Writing – original draft. **Yu-Dong Zhang:** Conceptualization, Methodology, Investigation, Writing – original draft. **Juan M. Górriz:** Conceptualization, Validation, Supervision, Investigation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] Y. Zheng, X. Hu, Concurrent prediction of finger forces based on source separation and classification of neuron discharge information, Int. J. Neural Syst. 31 (06) (2021) 2150010, http://dx.doi.org/10.1142/S0129065721500106.

[2] W.Y. Peh, J. Thomas, E. Bagheri, R. Chaudhari, S. Karia, R. Rathakrishnan, V. Saini, N. Shah, R. Srivastava, Y.-L. Tan, J. Dauwels, Multi-center validation study of automated classification of pathological slowing in adult scalp electroencephalograms via frequency features, Int. J. Neural Syst. 31 (06) (2021) 2150016, http://dx.doi.org/10.1142/S0129065721500167.

[3] Y. Xue, H. Zhu, F. Neri, A self-adaptive multi-objective feature selection approach for classification problems, Integr. Comput.-Aided Eng. (2021) http://dx.doi.org/10.3233/ICA-210664.

[4] J. Buenaposada, L. Baumela, Improving multi-class boosting-based object detection, Integr. Comput.-Aided Eng. 28 (2020) 1–16, http://dx.doi.org/10.3233/ICA-200636.

[5] S. Liapis, K. Christantonis, V. Chazan-Pantzalis, A. Manos, D. Filippidou, C. Tjortjis, A methodology using classification for traffic prediction: Featuring the impact of COVID-19, Integr. Comput.-Aided Eng. 28 (2021) 1–19, http://dx.doi.org/10.3233/ICA-210663.

[6] J. Górriz, J. Ramírez, A. Ortíz, F. Martínez-Murcia, F. Segovia, J. Suckling, M. Leming, Y. Zhang, J. Álvarez-Sánchez, G. Bologna, P. Bonomini, F. Casado, D. Charte, F. Charte, R. Contreras, A. Cuesta-Infante, A. Duro, A. Fernández-Caballero, J. Fernández, Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications, Neurocomputing 410 (2020) 237–270, http://dx.doi.org/10.1016/j.neucom.2020.05.078.

[7] J.E. Arco, J. Ramírez, J.M. Górriz, M. Ruz, Data fusion based on searchlight analysis for the prediction of Alzheimer's disease, Expert Syst. Appl. 185 (2021) 115549, http://dx.doi.org/10.1016/j.eswa.2021.115549.

[8] D. Castillo-Barnes, J. Ramírez, F. Segovia, F.J. Martínez-Murcia, D. Salas-Gonzalez, J.M. Górriz, Robust ensemble classification methodology for I123-Ioflupane SPECT images and multiple heterogeneous biomarkers in the diagnosis of Parkinson's disease, Front. Neuroinform. 12 (2018) 53, http://dx.doi.org/10.3389/fninf.2018.00053.

[9] J.E. Arco, A. Ortiz, D. Castillo-Barnes, J.M. Górriz, J. Ramírez, Quantifying inter-hemispheric differences in Parkinson's disease using siamese networks, in: J.M. Ferrández Vicente, J.R. Álvarez-Sánchez, F. de la Paz López, H. Adeli (Eds.), Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications, Springer International Publishing, Cham, 2022, pp. 156–165.

[10] Y.D. Zhang, S.C. Satapathy, L.Y. Zhu, J.M. Górriz, S.H. Wang, A seven-layer convolutional neural network for chest CT based COVID-19 diagnosis using stochastic pooling, IEEE Sens. J. (2020) 1, http://dx.doi.org/10.1109/JSEN.2020.3025855.

[11] S.-H. Wang, V.V. Govindaraj, J.M. Górriz, X. Zhang, Y.-D. Zhang, Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network, Inf. Fusion 67 (2021) 208–229, URL http://www.sciencedirect.com/science/article/pii/S1566253520303705.

[12] T.B. Chandra, K. Verma, B.K. Singh, D. Jain, S.S. Netam, Automatic detection of tuberculosis related abnormalities in Chest X-ray images using hierarchical feature extraction scheme, Expert Syst. Appl. 158 (2020) 113514, http://dx.doi.org/10.1016/j.eswa.2020.113514.

[13] M.A. Elaziz, K.M. Hosny, A. Salah, M.M. Darwish, S. Lu, A.T. Sahlol, New machine learning method for image-based diagnosis of COVID-19, PLoS One 15 (6) (2020) 1–18, http://dx.doi.org/10.1371/journal.pone.0235187.

[14] J.E. Arco, A. Ortiz, J. Ramírez, Y.-D. Zhang, J.M. Górriz, Tiled sparse coding in Eigenspaces for image classification, Int. J. Neural Syst. 32 (03) (2022) 2250007, http://dx.doi.org/10.1142/S0129065722500071.

[15] A. Lozano, J.S. Suárez, C. Soto-Sánchez, F.J. Garrigós, J.J. Martínez-Álvarez, J.M. Ferrández, E. Fernández, Neurolight: A deep learning neural interface for cortical visual prostheses, Int. J. Neural Syst. (2020) 2050045.

[16] P. Lara-Benítez, M. Carranza-García, J.C. Riquelme, An experimental review on deep learning architectures for time series forecasting, Int. J. Neural Syst. 31 (03) (2021) 2130001, http://dx.doi.org/10.1142/s0129065721300011.

[17] M.A. Ozdemir, O.K. Cura, A. Akan, Epileptic EEG classification by using time-frequency images for deep learning, Int. J. Neural Syst. 31 (08) (2021) 2150026, http://dx.doi.org/10.1142/S012906572150026X.

[18] P. Lara-Benítez, M. Carranza-García, J. García-Gutiérrez, J. Riquelme, Asynchronous dual-pipeline deep learning framework for online data stream classification, Integr. Comput.-Aided Eng. 27 (2020) 1–19, http://dx.doi.org/10.3233/ICA-200617.

[19] J. García-González, J. Ortiz-De-Lazcano-Lobato, R.M. Luque Baena, E. López-Rubio, Background subtraction by probabilistic modeling of patch features learned by deep autoencoders, Integr. Comput.-Aided Eng. 27 (2020) 1–13, http://dx.doi.org/10.3233/ICA-200621.

[20] S. Hamreras, B. Boucheham, M.A. Molina-Cabello, R. Benítez-Rochel, E. López-Rubio, Content based image retrieval by ensembles of deep learning object classifiers, Integr. Comput.-Aided Eng. 27 (2020) 1–15, http://dx.doi.org/10.3233/ICA-200625.

[21] Y. Hou, T. Chen, X. Lun, F. Wang, A novel method for classification of multi-class motor imagery tasks based on feature fusion, Neurosci. Res. (2021) http://dx.doi.org/10.1016/j.neures.2021.09.002.

[22] H. Noğay, H. Adeli, Detection of epileptic seizure using pretrained deep convolutional neural network and transfer learning, Eur. Neurol. 83 (2020) 602–614, http://dx.doi.org/10.1159/000512985.

[23] A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision? in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, pp. 5574–5584.

[24] Y. Gal, Uncertainty in Deep Learning (Ph.D. thesis), University of Cambridge, 2016.

[25] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, in: NIPS'12, 2012, pp. 1097–1105.

[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[27] M. Lin, Q. Chen, S. Yan, Network in network, 2014, CoRR abs/1312.4400.

[28] D.A. Borges Oliveira, L.G. Ribeiro Pereira, T. Bresolin, R.E. Pontes Ferreira, J.R. Rebouças Dorea, A review of deep learning algorithms for computer vision systems in livestock, Livest. Sci. 253 (2021) 104700, http://dx.doi.org/10.1016/j.livsci.2021.104700.

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[30] A. Veit, M. Wilber, S. Belongie, Residual networks behave like ensembles of relatively shallow networks, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 550–558.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-ResNet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 4278–4284.

[32] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807.

[33] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, B. Kim, Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognit. 120 (2021) 108102, http://dx.doi.org/10.1016/j.patcog.2021.108102.

[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929, http://dx.doi.org/10.1109/CVPR.2016.319.

[35] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626, http://dx.doi.org/10.1109/ICCV.2017.74.

[36] C. Barata, M.E. Celebi, J.S. Marques, Explainable skin lesion diagnosis using taxonomies, Pattern Recognit. 110 (2021) 107413, http://dx.doi.org/10.1016/j.patcog.2020.107413.

[37] R.C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3449–3457, http://dx.doi.org/10.1109/ICCV.2017.371.

[38] J. Wagner, J.M. Köhler, T. Gindele, L. Hetzel, J.T. Wiedemer, S. Behnke, Interpretable and fine-grained visual explanations for convolutional neural networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9089–9099.

[39] Y. Chai, Y. Bian, H. Liu, J. Li, J. Xu, Glaucoma diagnosis in the Chinese context: An uncertainty information-centric Bayesian deep learning model, Inf. Process. Manage. 58 (2) (2021) 102454, http://dx.doi.org/10.1016/j.ipm.2020.102454.

[40] T. Chavez, N. Vohra, K. Bailey, M. El-Shenawee, J. Wu, Supervised Bayesian learning for breast cancer detection in terahertz imaging, Biomed. Signal Process. Control 70 (2021) 102949, http://dx.doi.org/10.1016/j.bspc.2021.102949.

[41] H. Ma, I. Smal, J. Daemen, T. van Walsum, Dynamic coronary roadmapping via catheter tip tracking in X-ray fluoroscopy with deep learning based Bayesian filtering, Med. Image Anal. 61 (2020) 101634, http://dx.doi.org/10.1016/j.media.2020.101634.

[42] A. Garifullin, L. Lensu, H. Uusitalo, Deep Bayesian baseline for segmenting diabetic retinopathy lesions: Advances and challenges, Comput. Biol. Med. 136 (2021) 104725, http://dx.doi.org/10.1016/j.compbiomed.2021.104725.

[43] E.H. Houssein, M.M. Emam, A.A. Ali, P.N. Suganthan, Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review, Expert Syst. Appl. 167 (2021) 114161, http://dx.doi.org/10.1016/j.eswa.2020.114161, URL https://www.sciencedirect.com/science/article/pii/S0957417420309015.

[44] F. Martínez-Murcia, J. Gorriz, J. Ramírez, A. Ortiz, Convolutional neural networks for neuroimaging in Parkinson's disease: Is preprocessing needed? Int. J. Neural Syst. 28 (2018) http://dx.doi.org/10.1142/S0129065718500351.

[45] A. Ortiz, F. Martínez-Murcia, M. García-Tarifa, F. Lozano, J. Gorriz, J. Ramírez, Automated diagnosis of parkinsonian syndromes by deep sparse filtering-based features, in: Innovation in Medicine and Healthcare 2016, 2016, pp. 249–258, http://dx.doi.org/10.1007/978-3-319-39687-3_24.

[46] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, Neural Inf. Process. Syst. 25 (2012) http://dx.doi.org/10.1145/3065386.

[47] F. Segovia, M. García-Pérez, J. Gorriz, J. Ramírez, F. Martínez-Murcia, Assisting the diagnosis of neurodegenerative disorders using principal component analysis and TensorFlow, in: International Joint Conference SOCO'16-CISIS'16-ICEUTE'16, 2016, pp. 43–52, http://dx.doi.org/10.1007/978-3-319-47364-2_5.

[48] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. 61 (2015) 85–117, http://dx.doi.org/10.1016/j.neunet.2014.09.003, URL http://www.sciencedirect.com/science/article/pii/S0893608014002135.

[49] A. Payan, G. Montana, Predicting alzheimer's disease: a neuroimaging study with 3D convolutional neural networks, in: ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings, Vol. 2, 2015.

[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[51] H. Wang, D. Yeung, Towards Bayesian deep learning: A framework and some existing methods, IEEE Trans. Knowl. Data Eng. 28 (2016) 3395–3408.

[52] A.D. Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2) (2009) 105–112, http://dx.doi.org/10.1016/j.strusafe.2008.06.020, Risk Acceptance and Risk Communication.

[53] D.A. Nix, A.S. Weigend, Estimating the mean and variance of the target probability distribution, in: Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), Vol. 1, 1994, pp. 55–60, http://dx.doi.org/10.1109/ICNN.1994.374138, vol. 1.

[54] Q.V. Le, A.J. Smola, S. Canu, Heteroscedastic Gaussian process regression, in: Proceedings of the 22nd International Conference on Machine Learning, in: ICML '05, Association for Computing Machinery, 2005, pp. 489–496, http://dx.doi.org/10.1145/1102351.1102413.

[55] R. Cipolla, Y. Gal, A. Kendall, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7482–7491, http://dx.doi.org/10.1109/CVPR.2018.00781.

[56] A. Damianou, N. Lawrence, in: C.M. Carvalho, P. Ravikumar (Eds.), Deep Gaussian Processes, in: Proceedings of Machine Learning Research, vol. 31, PMLR, 2013, pp. 207–215.

[57] L.V. Jospin, W.L. Buntine, F. Boussaid, H. Laga, M. Bennamoun, Hands-on Bayesian neural networks - a tutorial for deep learning users, 2020, ArXiv abs/2007.06823.

[58] Y. Li, Y. Gal, Dropout inference in Bayesian neural networks with Alpha-divergences, in: ICML, 2017.

[59] T.B. Chandra, K. Verma, B.K. Singh, D. Jain, S.S. Netam, Coronavirus disease (COVID-19) detection in Chest X-Ray images using majority voting based classifier ensemble, Expert Syst. Appl. 165 (2021) 113909, http://dx.doi.org/10.1016/j.eswa.2020.113909.

[60] T. Zhou, H. ling Lu, Z. Yang, S. Qiu, B. qiang Huo, Y. Dong, The ensemble deep learning model for novel COVID-19 on CT images, Appl. Soft Comput. (2020) 106885, http://dx.doi.org/10.1016/j.asoc.2020.106885.

[61] M. Liu, D. Zhang, D. Shen, Ensemble sparse classification of Alzheimer's disease, NeuroImage 60 (2) (2012) 1106–1116, http://dx.doi.org/10.1016/j.neuroimage.2012.01.055.

[62] X. Gao, Y. He, M. Zhang, X. Diao, X. Jing, B. Ren, W. Ji, A multiclass classification using one-versus-all approach with the differential partition sampling ensemble, Eng. Appl. Artif. Intell. 97 (2021) 104034, http://dx.doi.org/10.1016/j.engappai.2020.104034.

[63] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes, Pattern Recognit. 44 (8) (2011) 1761–1776, http://dx.doi.org/10.1016/j.patcog.2011.01.017.

[64] L. Zhou, H. Fujita, Posterior probability based ensemble strategy using optimizing decision directed acyclic graph for multi-class classification, Inform. Sci. 400–401 (2017) 142–156, http://dx.doi.org/10.1016/j.ins.2017.02.059, URL http://www.sciencedirect.com/science/article/pii/S0020025516314207.

[65] Kaggle, Chest X-Ray images (pneumonia) dataset, 2020, URL https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia?.

[66] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M.K. Prasadha, J. Pei, M.Y. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V.A. Huu, C. Wen, E.D. Zhang, C.L. Zhang, O. Li, X. Wang, M.A. Singer, X. Sun, J. Xu, A. Tafreshi, M.A. Lewis, H. Xia, K. Zhang, Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122 – 1131.e9, http://dx.doi.org/10.1016/j.cell.2018.02.010.

[67] The Parkinson progression marker initiative (PPMI), Prog. Neurobiol. 95 (4) (2011) 629–635, http://dx.doi.org/10.1016/j.pneurobio.2011.09.005.

[68] J.C. Mazziotta, A.W. Toga, A. Evans, P.T. Fox, J.L. Lancaster, K. Zilles, R.P. Woods, T. Paus, G. Simpson, B. Pike, C.J. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N.J. Kabani, G.L. Goualher, D.I. Boomsma, T.D. Cannon, R. Kawashima, B. Mazoyer, A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM), Philos. Trans. R. Soc. Lond. Ser. B 356 1412 (2001) 1293–1322.

[69] A. Ortiz, J. Munilla, M. Martínez-Ibañez, J.M. Górriz, J. Ramírez, D. Salas-Gonzalez, Parkinson's disease detection using isosurfaces-based features and convolutional neural networks, Front. Neuroinform. 13 (2019) 48, http://dx.doi.org/10.3389/fninf.2019.00048.

[70] A. Ortiz, F.J. Martínez Murcia, J. Munilla, J.M. Górriz, J. Ramírez, Label aided deep ranking for the automatic diagnosis of Parkinsonian syndromes, Neurocomputing 330 (2019) 162–171, http://dx.doi.org/10.1016/j.neucom.2018.10.074.

[71] F. Segovia, J.M. Górriz, J. Ramírez, F.J. Martínez-Murcia, J. Levin, M. Schuberth, M. Brendel, A. Rominger, K. Bötzel, G. Garraux, C. Phillips, Multivariate analysis of 18f-DMFP PET data to assist the diagnosis of parkinsonism, Front. Neuroinform. 11 (2017) 23, http://dx.doi.org/10.3389/fninf.2017.00023.

[72] D. Castillo-Barnes, J. Ramírez, F. Segovia, F.J. Martínez-Murcia, D. Salas-Gonzalez, J.M. Górriz, Robust ensemble classification methodology for I123-ioflupane SPECT images and multiple heterogeneous biomarkers in the diagnosis of Parkinson's Disease, Front. Neuroinform. 12 (2018) 53, http://dx.doi.org/10.3389/fninf.2018.00053.

[73] I.A. Illán, J.M. Gorrz, J. Ramírez, F. Segovia, J.M. Jiménez-Hoyuela, S.J.O. Lozano, Automatic assistance to parkinson's disease diagnosis in DaTSCAN SPECT imaging, Med. Phys. 39 10 (2012) 5971–5980.

[74] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, in: IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 1137–1143.

[75] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, J. Thorac. Oncol. 5 (9) (2010) 1315–1316, http://dx.doi.org/10.1097/JTO.0b013e3181ec173d.

[76] K. Hajian-Tilaki, Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, Casp. J. Intern. Med. 4 (2013) 627–635.

[77] J.J. Rodriguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: A new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1619–1630, http://dx.doi.org/10.1109/TPAMI.2006.211.

[78] J. Wang, Y. Yang, B. Xia, A simplified cohen's kappa for use in binary classification data annotation tasks, IEEE Access 7 (2019) 164386–164397, http://dx.doi.org/10.1109/ACCESS.2019.2953104.

[79] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.

[80] B. Di Eugenio, M. Glass, The Kappa statistic: A second look, Comput. Linguist. 30 (1) (2004) 95–101, http://dx.doi.org/10.1162/089120104773633402.

[81] G. White, Basics of estimating measurement uncertainty, Clin. Biochem. Rev. / Aust. Assoc. Clin. Biochem. 29 (Suppl 1) (2008) S53–60.

[82] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017.

[83] S. Tabik, A. Gómez-Ríos, J.L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J.L. Suárez, J. Luengo, M.A. Valero-González, P. Garcí a Villanova, E. Olmedo-Sánchez, F. Herrera, COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-Ray images, IEEE J. Biomed. Health Inf. 24 (12) (2020) 3595–3605, http://dx.doi.org/10.1109/JBHI.2020.3037127.

[84] S.-H. Wang, D.R. Nayak, D.S. Guttery, X. Zhang, Y.-D. Zhang, COVID-19 classification by ccshnet with deep fusion using transfer learning and discriminant correlation analysis, Inf. Fusion 68 (2021) 131–148, http://dx.doi.org/10.1016/j.inffus.2020.11.005.

[85] G. Muhammad, M. Shamim Hossain, Covid-19 and non-COVID-19 classification using multi-layers fusion from lung ultrasound images, Inf. Fusion 72 (2021) 80–88, http://dx.doi.org/10.1016/j.inffus.2021.02.013.

[86] M.R. Ibrahim, S. Youssef, K.M. Fathalla, Abnormality detection and intelligent severity assessment of human chest computed tomography scans using deep learning: a case study on SARS-COV-2 assessment, J. Ambient Intell. Humaniz. Comput. (2021) 1–24.

[87] S.-H. Wang, V.V. Govindaraj, J.M. Górriz, X. Zhang, Y.-D. Zhang, Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network, Inf. Fusion 67 (2021) 208–229, http://dx.doi.org/10.1016/j.inffus.2020.10.004.

[88] A. Zargari Khuzani, M. Heidari, S.A. Shariati, COVID-classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images, Sci. Rep. 11 (1) (2021) 9887, http://dx.doi.org/10.1038/s41598-021-88807-2.

[89] W. Zhao, W. Jiang, X. Qiu, Deep learning for COVID-19 detection based on CT images, Sci. Rep. 11 (2021) 14353, http://dx.doi.org/10.1038/s41598-021-93832-2.

[90] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U. Rajendra Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images, Comput. Biol. Med. 121 (2020) 103792, http://dx.doi.org/10.1016/j.compbiomed.2020.103792.

[91] M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N.A. Emadi, M.B.I. Reaz, M.T. Islam, Can AI help in screening viral and COVID-19 Pneumonia? IEEE Access 8 (2020) 132665–132676, http://dx.doi.org/10.1109/ACCESS.2020.3010287.

[92] P. Kedia, Anjum, R. Katarya, CoVNet-19: A deep learning model for the detection and analysis of COVID-19 patients, Appl. Soft Comput. 104 (2021) 107184, http://dx.doi.org/10.1016/j.asoc.2021.107184.

[93] A. Gupta, Anjum, S. Gupta, R. Katarya, InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using chest X-ray, Appl. Soft Comput. 99 (2021) 106859, http://dx.doi.org/10.1016/j.asoc.2020.106859.

[94] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, X. Liu, Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays, Pattern Recognit. 110 (2021) 107613, http://dx.doi.org/10.1016/j.patcog.2020.107613.

[95] B. Nigam, A. Nigam, R. Jain, S. Dodia, N. Arora, B. Annappa, Covid-19: Automatic detection from X-ray images by utilizing deep learning methods, Expert Syst. Appl. 176 (2021) 114883, http://dx.doi.org/10.1016/j.eswa.2021.114883.

[96] A. Alhudhaif, K. Polat, O. Karaman, Determination of COVID-19 pneumonia based on generalized convolutional neural network model from chest X-ray images, Expert Syst. Appl. 180 (2021) 115141, http://dx.doi.org/10.1016/j.eswa.2021.115141.

[97] M. Shorfuzzaman, M.S. Hossain, MetaCOVID: A siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients, Pattern Recognit. 113 (2021) 107700, http://dx.doi.org/10.1016/j.patcog.2020.107700.

[98] A. Dixit, A. Mani, R. Bansal, CoV2-detect-net: Design of COVID-19 prediction model based on hybrid DE-PSO with SVM using chest X-ray images, Inform. Sci. (2021) http://dx.doi.org/10.1016/j.ins.2021.03.062.

[99] C. Li, Y. Yang, H. Liang, B. Wu, Transfer learning for establishment of recognition of COVID-19 on CT imaging using small-sized training datasets, Knowl.-Based Syst. 218 (2021) 106849, http://dx.doi.org/10.1016/j.knosys.2021.106849.

[100] T. Ozcan, A new composite approach for COVID-19 detection in X-ray images using deep features, Appl. Soft Comput. 111 (2021) 107669, http://dx.doi.org/10.1016/j.asoc.2021.107669.

[101] G. Marques, D. Agarwal, I. de la Torre Díez, Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network, Appl. Soft Comput. 96 (2020) 106691, http://dx.doi.org/10.1016/j.asoc.2020.106691.

[102] O. Cigdem, H. Demirel, D. Unay, The performance of local-learning based clustering feature selection method on the diagnosis of parkinson's disease using structural MRI, in: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 2019, pp. 1286–1291, http://dx.doi.org/10.1109/SMC.2019.8914611.

[103] Y. Wu, J.-H. Jiang, L. Chen, J.-Y. Lu, J.-J. Ge, F.-T. Liu, J.-T. Yu, W. Lin, C.-T. Zuo, J. Wang, Use of radiomic features and support vector machine to distinguish Parkinson's disease cases from normal controls, Ann. Transl. Med. 7 (23) (2019).

[104] H. Choi, S. Ha, H.-J. Im, S.H. Paek, D.S. Lee, Refining diagnosis of parkinson's disease with deep learning-based interpretation of dopamine transporter imaging, NeuroImage : Clin. 16 (2017) 586–594.

[105] P.R. Magesh, R.D. Myloth, R.J. Tom, An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery, Comput. Biol. Med. 126 (2020) 104041, http://dx.doi.org/10.1016/j.compbiomed.2020.104041.

[106] Y. Yang, L. Wei, Y. Hu, Y. Wu, L. Hu, S. Nie, Classification of parkinson's disease based on multi-modal features and stacking ensemble learning, J. Neurosci. Methods 350 (2021) 109019, http://dx.doi.org/10.1016/j.jneumeth.2020.109019.

[107] B. Gong, J. Shi, S. Ying, Y. Dai, Q. Zhang, Y. Dong, H. An, Y. Zhang, Neuroimaging-based diagnosis of Parkinson's disease with deep neural mapping large margin distribution machine, Neurocomputing 320 (2018) 141–149, http://dx.doi.org/10.1016/j.neucom.2018.09.025.

[108] C.-Y. Chien, S.-W. Hsu, T.-L. Lee, P.-S. Sung, C.-C. Lin, Using artificial neural network to discriminate Parkinson's Disease from other Parkinsonisms by focusing on Putamen of Dopamine Transporter SPECT images, Biomedicines 9 (1) (2021).

[109] A. Piccardo, R. Cappuccio, G. Bottoni, D. Cecchin, A. Mazzella, A. Cirone, S. Righi, M. Ugolini, P. Bianchi, P. Bertolaccini, E. Lorenzini, M. Massollo, A. Castaldi, F. Fiz, L. Strada, A. Cistaro, M. Sette, The role of the deep convolutional neural network as an aid to interpreting brain [18F]DOPA PET/CT in the diagnosis of Parkinson's disease, Eur. Radiol. 31 (2021) 7003–7011, http://dx.doi.org/10.1007/s00330-021-07779-z.

[110] S. Chakraborty, S. Aich, H.-C. Kim, Detection of Parkinson's disease from 3T T1 weighted MRI scans using 3D convolutional neural network, Diagnostics 10 (6) (2020) http://dx.doi.org/10.3390/diagnostics10060402.

[111] Y. Dai, Z. Tang, Y. Wang, Z. Xu, Data driven intelligent diagnostics for Parkinson's disease, IEEE Access 7 (2019) 106941–106950, http://dx.doi.org/10.1109/ACCESS.2019.2931744.

[112] S.-Y. Hsu, L.-R. Yeh, T.-B. Chen, W.-C. Du, Y.-H. Huang, W.-H. Twan, M.-C. Lin, Y.-H. Hsu, Y.-C. Wu, H.-Y. Chen, Classification of the multiple stages of parkinson's disease by a deep convolution neural network based on 99mTc-TRODAT-1 SPECT images, Molecules 25 (20) (2020) http://dx.doi.org/10.3390/molecules25204792.

[113] K. Yasaka, K. Kamagata, T. Ogawa, T. Hatano, H. Takeshige-Amano, K. Ogaki, C. Andica, H. Akai, A. Kunimatsu, W. Uchida, N. Hattori, S. Aoki, O. Abe, Parkinson's disease: deep learning with a parameter-weighted structural connectome matrix for diagnosis and neural circuit disorder investigation, Neuroradiology 63 (2021) 1–12, http://dx.doi.org/10.1007/s00234-021-02648-4.

[114] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, B. Xu, A deep learning algorithm using CT images to screen for corona virus disease (COVID-19), MedRxiv (2020) http://dx.doi.org/10.1101/2020.02.14.20023028.

[115] E.E.-D. Hemdan, M.A. Shouman, M.E. Karar, COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-Ray images, 2020, arXiv:2003.11055.

[116] I. Apostolopoulos, M. Tzani, Covid-19: Automatic detection from X-Ray images utilizing transfer learning with convolutional neural networks, Australas. Phys. Eng. Sci. Med. / Supported Australas. Coll. Phys. Sci. Med. Australas. Assoc. Phys. Sci. Med. 43 (2020) http://dx.doi.org/10.1007/s13246-020-00865-4.